

TRACKING PILOT INTERACTIONS WITH FLIGHT MANAGEMENT SYSTEMS THROUGH EYE MOVEMENTS

Melanie Diez, Deborah A. Boehm-Davis, Robert W. Holt, Mary E. Pinney, Jeffrey T. Hansberger, and
Wolfgang Schoppek
George Mason University
Fairfax, VA

ABSTRACT

Although automation has benefits for commercial aviation, it has led to undesirable consequences. One approach to understanding errors is the development and examination of cognitive models of the flying task. However, the construction of these models requires knowledge about the processes pilots use when they fly and how they acquire readings from their flight instruments. We explored this issue by collecting data from pilots interacting with a Boeing 747-400 desktop simulator. Eye track data provided information about where pilots were looking. This report describes the data obtained and provides suggestions for what these data mean in light of cognitive models.

INTRODUCTION

Despite many benefits (Wiener, 1988), aircraft automation has led to several undesirable consequences. The generally high reliability of automated systems has led to the possibility of automation-induced complacency (Parasuraman, Molloy, & Singh, 1993). Complacency can lead to decreased monitoring of the system and a decreased likelihood of detecting system malfunctions. Furthermore, not experiencing malfunctions during normal operations may decrease the crew's skill or facility in dealing with automation failures when they occur. Finally, some automation has been poorly designed and operates in a "clumsy" manner that may increase rather than decrease crew workload (Wiener, 1988).

The increased complexity and autonomy of advanced automation may lead to an increasing likelihood that the human crew is not aware of the current state of the automation and cannot correctly anticipate its future actions. A lack of mode awareness can cause the crew to take inappropriate actions or delay taking appropriate actions until it is too late (Sarter & Woods, 2000). Autonomous automation actions such as uncommanded mode changes may result in automation surprises that can distract the crew's attention from flight-critical tasks or otherwise adversely affect crew performance (Sarter & Woods, 1994, 1995, 1997).

One possible approach to reducing these problems is to change the design, interface, or operating characteristics of the automation. This is a good long-term strategy, but it does not help with the current operational use of automation. Approaches that could be implemented immediately in the operational context include changing the operating policies and procedures for the automation or increasing pilot knowledge and expertise concerning automation.

However, implementing any intervention to reduce automation problems requires an accurate representation of pilot and crew interactions with the automation. The design of any effective intervention must efficiently solve the problem while not causing unanticipated negative side effects. Therefore, the design of interventions must be based on knowledge of the pilot-automation interaction process that is both accurate and precise.

One approach to developing accurate knowledge of this process is to model pilot cognitions as they interact with the automation. These models can be constructed at the level of a cognitive task analysis (e.g., see Irving, Polson & Irving, 1994) or at the more detailed level of a computational cognitive model (e.g., see Doane & Sohn, 2000; Jones, Laird, Nielsen, Coulter, Kenny & Koss, 1999). Both modeling approaches track the cycle of events from the pilots' acquisition of information from flight displays to associated cognitive processing and appropriate actions. Since the pilot actions change the flight situation, these models must be able to describe dynamic changes in the pilot-aircraft system at a detailed level. Expressing these models in a computational cognitive architecture allows both the analysis of the course of typical automation problems and the exploration of the effectiveness of possible solutions.

To develop cognitive models of automation use, the information from a cognitive task analysis should be combined with information from the performance of pilots using the automation in typical flight scenarios. Determining what information is attended

to by the pilot is particularly important to model the first step in the cognitive cycle.

Although previous work has described pilot scanning behavior, most of these studies focused on scan patterns as a function of workload or expertise. These studies suggest that workload increases dwell time (Bunecke, 1987; Ephrath, Tole, Stephens, & Young, 1980) and that expert pilots have shorter dwells and more frequent glances than novices (Bellenkes, Wickens, & Kramer, 1997; Miller, 1973). However, these studies do not address how the information is encoded or how often it is used, particularly in automated aircraft. Further, they do not address the relationship between visual behavior and information processing. Both of these issues are critical in designing a model that mimics human performance.

This study used eye-tracking methodology to examine the information attended to by commercial pilots using automation to fly a desktop simulator. Because this research was exploratory, the hypotheses were general. First, we expected that there would be some patterns of consistency in the use of automation during certain phases of flight although we also expected individual differences. Second, we expected that attention directed to information on the automation display would predict the short-term retention of relevant information.

METHOD

Participants

Five commercial pilots, currently certified to fly the Boeing 777 aircraft and domiciled in the Washington, D.C. area participated in this study. Four of the pilots serve as first officers on the 777; one serves as a captain. Four of the pilots were male and one was female.

Simulator

The study used an Aerowinx PS1 desktop simulator. This simulator operates on a PC platform and, according to subject matter experts, provides a realistic representation of the Boeing 747-400 flight management system. Although the simulator uses keystroke and mouse input rather than realistic controls (e.g., stick or rudder controls), the automation software displays all relevant information available in the real aircraft and has all relevant automation functions operating in the same manner. Our pilots reported that the simulation functioned reasonably well in simulating the responses of a real aircraft.

Procedure

Participants were asked to fly two scenarios using the simulator. Scenario 1 involved take-off, climb, cruise, descent and approach phases of a flight and lasted approximately 50 minutes. Scenario 2 involved the descent and approach phases of a flight and lasted approximately 30 minutes. Order of the scenarios was counterbalanced; three of the pilots flew scenario 1 first while two flew scenario 2 first. The session took approximately two hours overall; the procedure included:

- 1) an orientation to the simulator (~15 min)
- 2) eye-tracker calibration (~5-10 minutes)
- 3) the first scenario (between 30-50 min)
- 4) break (10-15 min)
- 5) the second scenario (between 30-50 min)
- 6) post-session interview (5-10 min)

The pilot flew the simulator alone (i.e., without a co-pilot) while eye movements were recorded using an ASL 504 eye tracker. Each session was conducted by an eye-tracker operator and a confederate pilot. The eye tracker operator calibrated the tracker and asked questions during the probe sessions. The confederate pilot played tape-recorded ATC clearances at the appropriate time and answered questions about the aviation situation. Eye-scan patterns, video, verbal protocol, and keystroke data were collected.

In addition, pilots were interrupted six times throughout the scenarios and asked to a) recall as many details of the current flying situation as possible, and b) recall values from specified instruments. These interruptions are referred to as “free recall” and “cued recall” respectively. During the free recall sessions, comments ranged from actual values of flight parameters to vague estimates. Responses often included comments on their own performance and/or the simulator’s performance.

During the cued recall sessions, pilots were asked about the values of specific flight parameters (altitude, heading, airspeed, MCP altitude, MCP heading, MCP airspeed, aircraft position, power setting, autothrottle mode, pitch mode, roll mode, autopilot and flight directors status.) The parameters probed never changed from session to session; however, cues were not probed in the cued recall session if the value already had been given during the free recall session.

After pilots became familiar with the structure of the cued recall sessions, they often anticipated some of these questions by mentioning them in the free recall. On the other hand, some values were never recalled correctly, even though the pilots knew they

would be asked for them. Furthermore, pilots consistently included qualitative information that we never asked for, such as wind speed and direction.

RESULTS

Eye Scan Data

The computer display of the cockpit was segmented into "areas of interest" (AOI) that represented instruments pilots might examine during flight (e.g., the mode control panel or the primary flight display). The mean fixation duration for each AOI was calculated for each of the four pilots with good eye data. Overall, mean durations ranged from a high of approximately 700 msec to a low of approximately 300 msec. Interestingly, the durations showed more variability across pilots for instruments that require manual interactions (e.g., mode control panel, control display unit). This could be due to pilots monitoring the values as they input them or to difficulties in manipulating the simulator interface. Conversely, mean fixation time was more consistent for instruments that did not require interaction, such as the primary flight display (PFD), navigational display (ND), and engine indicating crew alerting system (EICAS).

The percent of time that pilots spent looking at each AOI was also calculated. These data suggested that pilots spent the majority of their time looking at four instruments -- the ND, the PFD, the mode control panel (MCP), and the control display unit (CDU). The time spent looking at each of these instruments can be seen in Figure 1. In the figure, flight segments a through e represent the five segments in scenario 1. These segments were created by four probe sessions (cued and free recall) that interrupted the scenario. Segments f through h represent the three segments in scenario 2 created by the two probe sessions.

The extent to which the pilots are consistent in the amount of time they spent studying the instruments was calculated using two measures of inter-rater reliability, correlation and systematic differences. For each of the AOI except the mode control panel, the pilots demonstrated reasonable inter-pilot correlations across the eight flight segments (see Table 1). The correlations among the fixation profiles were highest for the CDU. Although none of the correlations reached acceptable levels of significance, it must be remembered that the number of flight segments flown by each pilot was quite small (eight).

We also found systematic differences in the percent of time spent looking at a particular display across the eight flight segments. These data

suggested some interesting differences among the pilots. For example, pilot 1 spent less time than average studying the PFD ($t(6) = -3.21, p < .01$), but spent more time than average studying the ND ($t(6) = 3.50, p < .01$). For the MCP, pilot 5 spent more time than average ($t(6) = 3.47, p < .01$) while pilot 2 spent less time than average ($t(6) = -3.37, p < .01$) studying this instrument. This pattern was reversed for the CDU, where pilot 2 spent more time than average ($t(6) = 3.33, p < .01$) and pilot 5 spent less time ($t(6) = -2.45, p < .05$).

A replay of the flight (using the video captured during the experiment) suggests reasons for some of these differences. For example, pilot 5 spent more time overall in the PFD and less in the ND. This arises from segment c of scenario 1, where the videotape shows that this pilot commented on readings suggesting that the wings were "rocking". He then attempted to determine if those readings were indicating turbulence. As a result of focusing his attention on the PFD, he did not examine the ND as much. None of the other pilots detected any anomalies during this segment; in fact, no such anomalies were programmed into the scenario.

Glance Duration and Accuracy of Report

Another question concerns the extent to which study of the instruments leads to more accurate knowledge about the state of the aircraft. To address this question, we evaluated the relationship between the time spent studying individual instruments and the ability of each pilot to accurately report the value of the instrument during the cued recall. Unfortunately, sufficient data for this analysis was only available from pilots 3 and 5.

Response accuracy was classified in one of three categories: incorrect, close, or correct. These categories were developed on the basis of the type of response necessary. If words were required to answer the question (e.g., pitch mode), the response was "correct" if the pilot reported the exact words or abbreviations that were displayed on the PFD. The response was "close" if one word was equivalent to the actual status and one was not equivalent (e.g., VNAV PTH vs. VNAV SPD). If they did not say either of the words or abbreviations accurately, the response was coded as "incorrect". When a numerical response was required (e.g., airspeed), the difference between actual and perceived status was calculated and the differences were placed in descending order. The data were then separated into three approximately equivalent categories representing "correct", "close" and "incorrect" responses. However, the categories did not always contain exactly the same number of responses and

differed across the two pilots. There were some cases where more than a third of the responses were objectively correct (i.e., the difference between the actual and perceived status was zero).

The average fixation time on the relevant instrument was calculated for correct, incorrect, and close responses and can be seen in Figure 2. For pilot 5, an analysis of variance confirmed a significant difference among the three categories ($F(2,33) = 3.40, p < .05$). For this participant, planned comparisons using a one-tailed t -test showed that fixation duration was longer for correct than for close responses ($t(25) = 1.85, p < .05$), and for correct than for incorrect responses ($t(28) = 2.22, p < .05$). For pilot 3, the overall analysis of variance was not significant ($F(2,33) = 1.44, p > .05$); however, a planned comparison showed that fixation durations were longer for close responses ($t(13) = 1.80, p < .05$) than for incorrect responses.

Cued Recall

In general, pilots were quite good at remembering current altitude and air speed. Approximate position was also usually recalled, often by reference to a nearby waypoint. Pilots were also rather good estimators of engine power. Conversely, pilots were poor at describing their current heading quantitatively. They defended this by saying that they were aware of their general heading and it was unnecessary to know the exact heading. One pilot said that as long as the heading “bugs” lined up, they knew they were on track. Pilots also had difficulty in recalling the flight mode annunciations for auto throttle and pitch mode. Indeed, it seems as if the pilots only maintained a binary coding of the modes (i.e., either something is engaged or it is not). Specifically, they did not seem to know the variations of VNAV (e.g., VNAV PTH, VNAV SPD). In the worst cases, the pilot was unable to appropriately report mode terminology. Roll mode was the exception, although one could argue that this was the ‘easiest’ to remember since it was almost always set to LNAV or HDG SEL and did not change as often or as rapidly as the pitch or power modes.

DISCUSSION

Scan Patterns

The data from this study examined what instrumentation is studied by pilots in an automated cockpit. The scan data suggest that pilots are relatively consistent in terms of which instruments they rely on while flying using automated systems. Although the inter-pilot correlations were not statistically significant, this appeared to be due to having only eight flight segments as the basis for the

correlations. In fact, the average correlations were quite respectable in size for some of the instruments (range of .40 to .47) given the small number of observations.

Where systematic differences were noted, a review of the flight suggested that environmental circumstances contributed greatly to the departures from the average performance. However, not all systematic differences could be explained by the record, and these residual differences may indicate real inter-pilot differences in the typical use of automation.

Implications for Cognitive Modeling

The positive correlations among pilots combined with the significant individual differences suggest that our cognitive model needs to be flexible enough to adopt several scan strategies. For example, pilots were relatively consistent in the total amount of time they spent within the PFD but not as consistent within the ND during most flight segments. Interestingly, the overall mean fixation duration within both PFD and ND were similar -- approximately 400 msec. This suggests that the primary cause of variability in the amount of time spent in an AOI is the frequency rather than the duration of fixations.

Furthermore, the relationship between fixation duration and recall accuracy suggests that there are two types of fixations that we may need to model -- those that serve to monitor instruments and those that serve to acquire information from the instruments for further processing. The monitoring fixations might be classified as glances which do not involve in-depth processing. In contrast, the information processing fixations may take an additional 200-300 msec, but result in better working memory for the information.

Caveats

There were some concerns with simulator fidelity on the part of the pilots. There was a consensus among the pilots that there should have been a longer practice session including a take off and initial turn so that they could become familiar with the deviations from an actual 747-400. Another concern was the fact that one pilot was doing the work of a two-person crew. Consequently, it is not possible to draw strong conclusions regarding eye scan patterns in crew situations. Pilots also mentioned that the number of clearances issued was high. Combined with the single pilot operation, this made it frustrating for the pilots. On the other hand, some typical distracters on the line (e.g., flight

attendants knocking on the door and other chatter on the radio) were not present in this simulation.

Summary

Despite some concerns about the use of a desktop simulator and the use of a single pilot crew, the data from this study provided baseline information on which instruments pilots study while flying scenarios using automated systems. The data suggest some regularities in pilot scanning behavior as well as some differences. The percent of time spent looking at each cluster of instruments depended on the phase of flight and on specific environmental events. These data suggest that it would not be unreasonable to build initial models using an "average" model of pilot scanning behavior which are then adjusted for individual differences. Further, they suggest that it will be important to incorporate information on phase of flight and the operational context into future modeling efforts.

ACKNOWLEDGEMENTS

This research was supported by grant NAG 2-1289 from NASA and 99-G-010 from the FAA. The authors thank Philip Ikomi for help with data collection.

REFERENCES

- Bellenkes, A. H., Wickens, C.D. & Kramer, A.F. (1997). Visual scanning and pilot expertise: The role of attentional flexibility and mental model development. *Aviation Space and Environmental Medicine*, 68, 569-579.
- Bunecke, J. L. (1987). Quantifying some information processing requirements of the pilot's instrument crosscheck. *Proceedings of the Human Factors Society Annual Meeting*, 31, 1301-1305.
- Doane, S. M. & Sohn, Y. W. (2000). ADAPT: A predictive cognitive model of user visual attention and action planning. *User Modeling and User-Adapted Interaction*, 10, 1-45.
- Ephrath, A. R., Tole, J.R., Stephens, A. T., & Young, L. R. (1980). Instrument scan – Is it an indicator of the pilot's workload? *Proceedings of the Human Factors Society Annual Meeting*, 24, 257-258.
- Irving, S., Polson, P., & Irving, J.E. (1994). Applications of formal models of human-computer interaction to training and use of the control and display unit. *Technical Report #94-08*. Boulder, CO: University of Colorado.
- Jones, R. M., Laird, J. E., Nielsen, P. E., Coulter, K. J., Kenny, P., & Koss, F. V. (1999). Automated intelligent pilots for combat flight simulation. *AI Magazine*, 27-41.
- Miller, J. M. (1973). Visual behavior changes of student pilots flying instrument approaches. *Proceedings of the Human Factors Society Annual Meeting*, 17, 208-214.
- Parasuraman, R., Molloy, R., & Singh, I.L. (1993) Performance consequences of automation-induced 'complacency', *International Journal of Aviation Psychology*, 3, 1-23.
- Sarter, N.D. & Woods, D. D. (1994) Pilot interaction with cockpit automation: II. An experimental study of pilots' model and awareness of the flight management and guidance system. *International Journal of Aviation Psychology*, 2, 303-321.
- Sarter, N.D. & Woods, D. D. (1995) How in the world did we ever get into that mode? *Human Factors*, 37, 5-19.
- Sarter, N.D. & Woods, D. D. (1997) Team play with a powerful and independent agent: a corpus of operational experiences and automation surprises on the airbus A-320. *Human Factors*, 39, 553-569.
- Sarter, N.D. & Woods, D. D. (2000) Team play with a powerful and independent agent: a full-mission simulation study. *Human Factors*, 42, 390-401.
- Wiener, E.L. (1988) Cockpit automation, in E. L. Wiener and D. C. Nagel (eds), *Human Factors in Aviation*. San Diego, California: Academic Press, 433-461.

Table 1. Consistency (as measured by correlation coefficients) between each pilot's percent of time spent looking at a particular display and the average time spent looking at a particular display across eight flight segments.

	Average	Pilot 1	Pilot 2	Pilot 3	Pilot 5
Primary Flight Display	.42	.55	.54	.21	.35
Navigational Display	.40	.58	.22	.23	.54
Mode Control Panel	.10	.10	-.18	.31	.17
Control Display Unit	.47	.25	.62	.41	.54

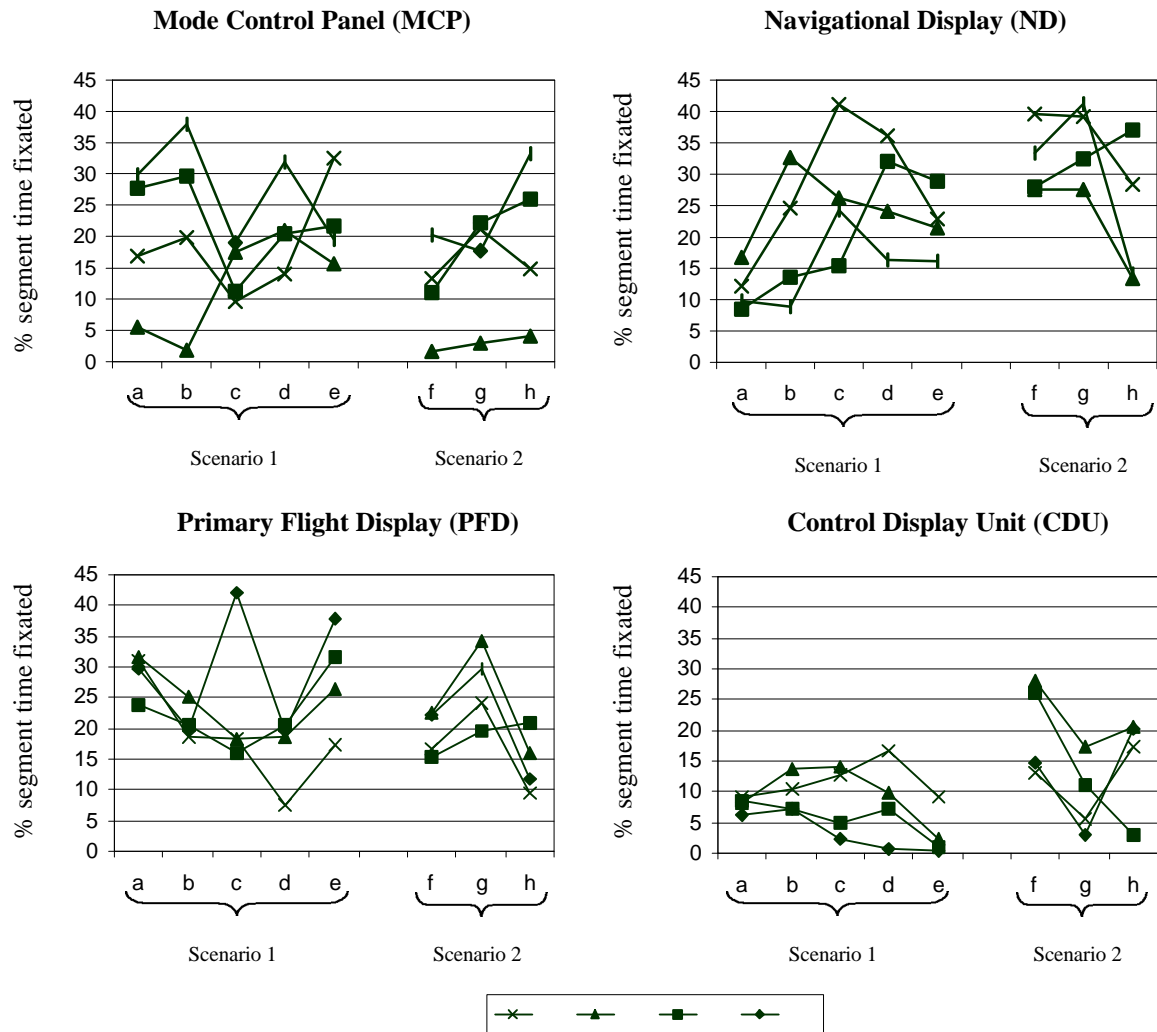


Figure 1. Percent of time (within each flight segment) spent in MCP, ND, PFD, and CDU

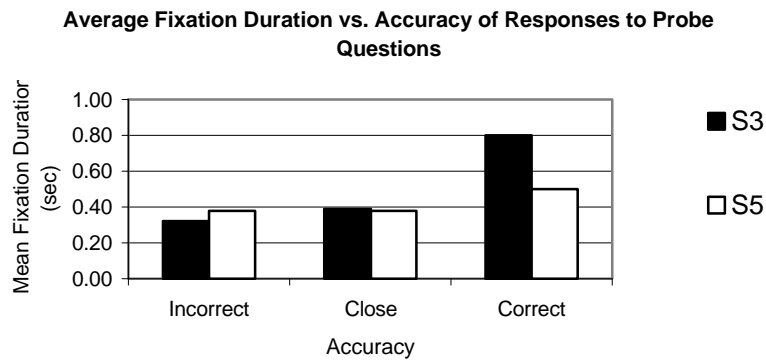


Figure 2. Data on fixation duration and correctness of response